THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút): https://www.youtube.com/watch?v=3lse2_0KY2o
- Link slides (dạng .pdf đặt trên Github của nhóm): https://github.com/TuananhSR/CS2205.CH201/blob/main/Anh%20Ng%C3%B4%20Tr%E1%BA%A7n%20Tu%E1%BA%A5n%20-%20CS2205.SEP2025.DeCuong.FinalReport.Template.Slide.pdf
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*
- *Lớp Cao học, mỗi nhóm một thành viên*

- Họ và Tên: Ngô Trần Tuấn Anh
- MSSV: 250101003

- Lớp: CS2205.CH201
- Tự đánh giá (điểm tổng kết môn): 9.5/10
- Số buổi vắng: 0
- Số câu hỏi QT cá nhân: 7
- Số câu hỏi QT của cả nhóm: 7
- Link Github: https://github.com/TuananhSR/CS2205.CH201

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

HƯỚNG TỚI MÃ HÓA HÌNH ẢNH CHÍNH XÁC: CẢI THIỆN TẠO ẢNH TỰ HỒI QUY VỚI LƯỢNG TỬ HÓA VECTOR ĐỘNG

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

TOWARDS ACCURATE IMAGE CODING: IMPROVED AUTOREGRESSIVE IMAGE GENERATION WITH DYNAMIC VECTOR QUANTIZATION

TÓM TẮT *(Tối đa 400 từ)*

Các mô hình sinh ảnh tự hồi quy (AR) dựa trên lượng tử hóa vector (VQ) hiện nay thường sử dụng mã hóa độ dài cố định, gây ra hiện tượng thiếu hụt chi tiết ở vùng phức tạp và dư thừa ở vùng đơn giản. Hạn chế này không chỉ làm giảm chất lượng hình ảnh mà còn dẫn đến thứ tự sinh ảnh raster-scan thiếu tự nhiên và lãng phí tài nguyên tính toán.

Để giải quyết vấn đề trên, nghiên cứu đề xuất một khung làm việc hai giai đoạn mới:

1. **Dynamic Quantization VAE (DQ-VAE):** Thực hiện mã hóa độ dài biến thiên thông qua mô-đun Mã hóa hạt động (DGC) và hàm Budget Loss. Cơ chế này giúp tự động phân bổ lượng mã tối ưu theo mật độ thông tin thực tế, đảm bảo biểu diễn hình ảnh chính xác và nhỏ gọn.
2. **DQ-Transformer:** Sinh ảnh theo lộ trình từ thô đến mịn (coarse-to-fine) dựa trên kiến trúc Stacked Transformer. Mô hình thực hiện dự đoán luân phiên vị trí và nội dung mã, giúp ưu tiên cấu trúc tổng thể trước khi hoàn thiện các chi tiết cục bộ, từ đó nâng cao tính nhất quán và chân thực.

Dự án sẽ tiến hành thực nghiệm trên các tập dữ liệu tiêu chuẩn như FFHQ và ImageNet. Mục tiêu trọng tâm là vượt mốc SOTA về chất lượng hình ảnh (chỉ số FID) và tối ưu hóa đáng kể tốc độ suy luận. Kết quả nghiên cứu kỳ vọng sẽ khẳng định mã hóa dựa trên mật độ thông tin động là giải pháp then chốt cho các hệ thống sinh ảnh hiệu suất cao thế hệ mới.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Trong kỷ nguyên sáng tạo nội dung số bằng AI, các mô hình tự hồi quy dựa trên định lượng vector (VQ) đóng vai trò nền tảng. Tuy nhiên, các phương pháp hiện hành (SOTA) gặp hạn chế lớn do mã hóa theo lưới cố định, phân bổ tài nguyên đồng đều bất kể độ phức tạp của vùng ảnh. Điều này dẫn đến một bài toán tính toán cần tối ưu:

- **Vấn đề tính toán (Computational Problem):** Sự mâu thuẫn giữa cấu trúc mã hóa tĩnh (fixed-length) và bản chất phân bổ thông tin không đồng đều của ảnh. Hệ quả là gây thiếu hụt chi tiết ở các vùng phức tạp và dư thừa tài nguyên ở các vùng đơn giản, làm giảm hiệu suất sinh ảnh và lãng phí bộ nhớ.
- **Lý do chọn đề tài và Tính thời sự:** Tối ưu hóa tốc độ và độ chính xác của mô hình sinh (Generative Models) là xu hướng then chốt. Đề tài trực tiếp giải quyết hạn chế của quét raster-scan bằng cách tiếp cận tự nhiên: mã hóa dựa trên mật độ thông tin.

Để giải quyết bài toán trên, chúng tôi đề xuất hệ thống mã hóa hình ảnh động gồm:

- **DQ-VAE:** Thực hiện mã hóa độ dài biến thiên (variable-length coding), tự động ưu tiên tài nguyên cho các vùng quan trọng để tái tạo chi tiết sắc nét.
- **DQ-Transformer:** Sinh ảnh theo lộ trình từ thô đến mịn (coarse-to-fine). Mô hình dự đoán khung xương ảnh trước khi hoàn thiện chi tiết phức tạp, giúp tăng tốc độ xử lý tổng thể.

Mô tả Input/Output:

- **Đầu vào (Input):** Hình ảnh tự nhiên (FFHQ, ImageNet) hoặc nhãn lớp (class labels).
- **Đầu ra (Output):** Hình ảnh tổng hợp chất lượng cao, đạt sự cân bằng giữa chỉ số FID (chất lượng) và FPS (tốc độ).

Khả năng ứng dụng thực tế:

- Nghiên cứu đóng góp trực tiếp vào việc xây dựng công cụ thiết kế đồ họa thông minh, nén dữ liệu thế hệ mới và tối ưu hóa AI cho các thiết bị cá nhân có tài nguyên hạn chế. Bằng việc chuyển đổi sang tư duy "mã hóa nhận thức nội dung", đề tài kỳ vọng thiết lập tiêu chuẩn mới trong lĩnh vực Image Coding.

MỤC TIÊU *(Viết trong vòng 3 mục tiêu)*

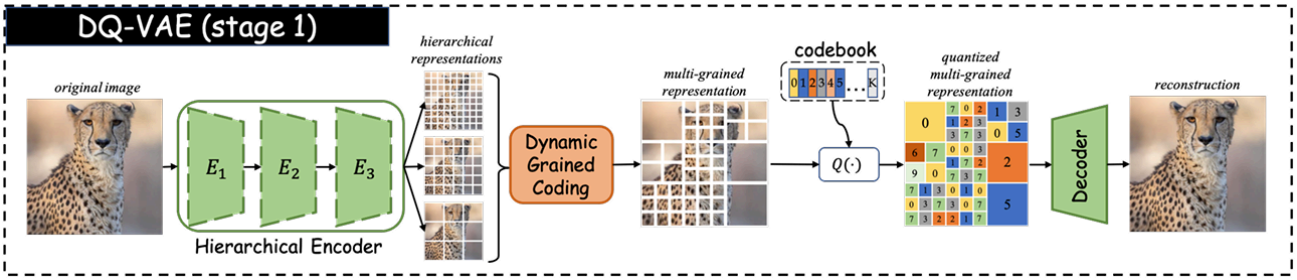
Để giải quyết các hạn chế về mã hóa cố định và thứ tự sinh ảnh không tự nhiên, nghiên cứu xác định ba mục tiêu trọng tâm sau:

1. **Xây dựng cơ chế mã hóa DQ-VAE dựa trên mật độ thông tin:** Phát triển mô-đun Dynamic Grained Coding (DGC) và hàm Budget Loss để tự động gán độ dài mã biến thiên. Mục tiêu là ưu tiên tài nguyên cho vùng phức tạp, đạt sự cân bằng tối ưu giữa độ chính xác tái tạo và tính nhỏ gọn của bảng mã.
2. **Thiết kế kiến trúc sinh ảnh phân cấp DQ-Transformer:** Xây dựng quy trình tạo ảnh từ thô đến mịn (coarse-to-fine) thông qua cấu trúc Stacked Transformer. Ứng dụng cơ chế dự đoán luân phiên vị trí và nội dung mã để thay thế quét raster-scan, đảm bảo tính nhất quán giữa cấu trúc tổng thể và chi tiết cục bộ.
3. **Đánh giá và tối ưu hiệu năng bằng thực nghiệm định lượng:** Kiểm chứng hệ thống trên bộ dữ liệu FFHQ và ImageNet nhằm chứng minh ưu thế vượt trội của giải pháp. Mục tiêu cụ thể là đạt mức cải thiện chỉ số FID từ 5-8% và gia tăng đáng kể tốc độ suy luận (Inference speed) so với các mô hình SOTA hiện nay.

NỘI DUNG VÀ PHƯƠNG PHÁP

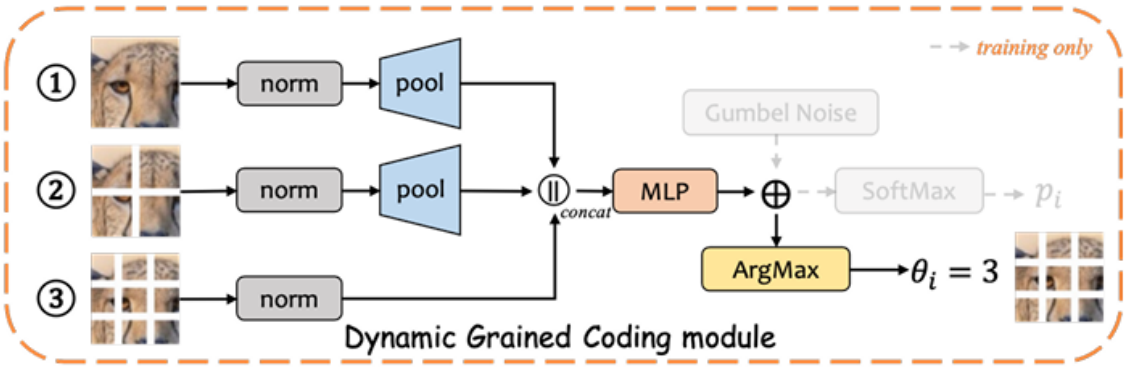
Để đạt được các mục tiêu đã đề ra, nghiên cứu dự kiến triển khai các nội dung và phương pháp kỹ thuật cụ thể theo ba giai đoạn chính sau đây:

1. Nghiên cứu và xây dựng hệ thống mã hóa hình ảnh động (DQ-VAE)



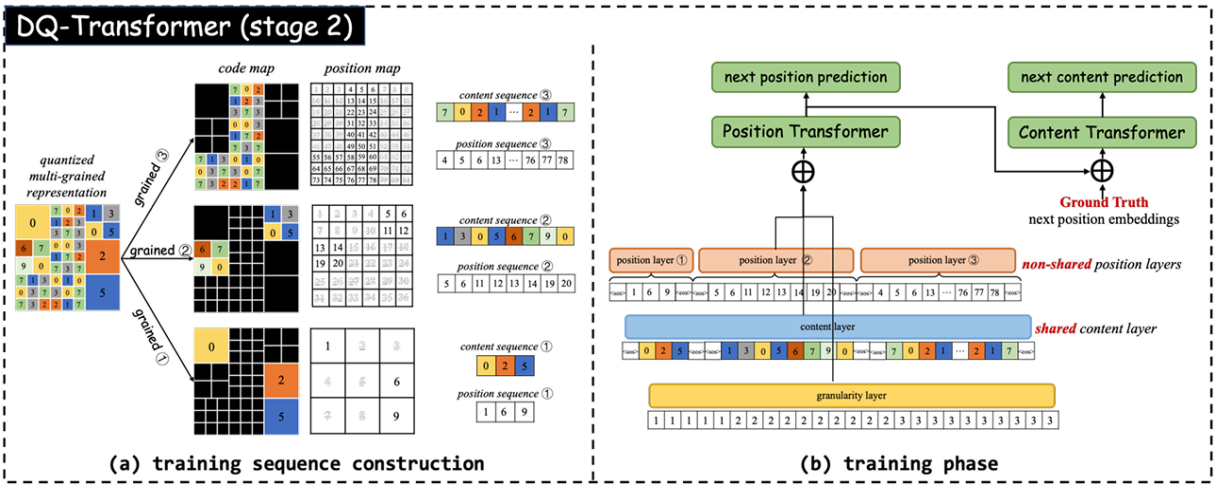
Hình 1: DQ-VAE gán mã có độ dài biến thiên cho từng vùng ảnh thông qua mô-đun Mã hóa Phân hạt Động (DGC).

- Nội dung: Phát triển cơ chế mã hóa có khả năng tự động điều chỉnh độ dài mã dựa trên mật độ thông tin thực tế của từng vùng ảnh, nhằm giải quyết sự thiếu hụt chi tiết ở các vùng phức tạp và sự dư thừa ở các vùng mịn.
- Phương pháp:
    - **Hierarchical Encoder:** Trích xuất đặc trưng hình ảnh tại nhiều cấp độ chi tiết (granularities) khác nhau.
    - **Mô-đun DGC (Dynamic Grained Coding):** Sử dụng gating network và kỹ thuật Gumbel-Softmax để lựa chọn cấp độ mã hóa linh hoạt, đảm bảo khả năng tính đạo hàm trong huấn luyện.
    - **Budget Loss:** Điều phối tỷ lệ phân bổ mã giữa các phân cấp, cân bằng tối ưu giữa chất lượng tái cấu trúc và tính nhỏ gọn của chuỗi mã.



Hình 2: Minh họa mô-đun Mã hóa Phân hạt Động.

2. Thiết kế mô hình tạo ảnh tự hồi quy phân cấp (DQ-Transformer)



Hình 3: DQ-Transformer mô hình hóa luân phiên vị trí và nội dung mã bằng các lớp Transformer xếp chồng, tạo ảnh tự hồi quy từ thô đến tinh.

- Nội dung: Xây dựng kiến trúc tạo ảnh theo trình tự từ thô đến mịn (coarse-to-fine), ưu tiên hình thành cấu trúc tổng thể trước khi hoàn thiện các chi tiết cục bộ, thay thế cho thứ tự quét raster-scan truyền thống.
- Phương pháp:
    - **Stacked Transformer:** Dự đoán luân phiên giữa Position-Transformer (vị trí mã) và Content-Transformer (nội dung mã).
    - **Lớp đầu vào chuyên biệt:** Kết hợp nhúng nội dung chung (shared-content) và vị trí riêng biệt (non-shared-position) để phân hóa chính xác vai trò mã giữa các phân cấp.
    - **Trình tự Coarse-to-fine:** Kiến tạo khung sườn từ vùng mịn (ít mã) trước, sau đó lấp đầy chi tiết từ vùng phức tạp (nhiều mã) để hoàn thiện hình ảnh tự nhiên.

3. Thực nghiệm, đánh giá và so sánh hiệu năng

- Nội dung: Tiến hành huấn luyện mô hình trên các tập dữ liệu quy mô lớn và thực hiện các phép đo kiểm chứng để xác định giá trị khoa học của giải pháp đề xuất.
- Phương pháp:
    - Thử nghiệm trên FFHQ và ImageNet cho các tác vụ sinh ảnh có và

không điều kiện.

- Định lượng qua FID, IS và tốc độ suy luận (Inference speed) để so sánh trực tiếp với các mô hình SOTA (ViT-VQGAN, RQ-VAE).
- Ablation Study nhằm kiểm chứng vai trò của mô-đun DGC, hàm Budget Loss và trình tự sinh ảnh coarse-to-fine đối với hiệu quả tổng thể.

KẾT QUẢ MONG ĐỢI

- Hiệu quả mã hóa (DQ-VAE): Chỉ số rFID dự kiến giảm 10-15% so với VQGAN; tái tạo chính xác các vùng mật độ thông tin cao.
- Chất lượng sinh ảnh (DQ-Transformer):
- FFHQ: FID đạt ngưỡng < 5.0 (cải thiện 5-8% so với ViT-VQGAN).
- ImageNet: Inception Score (IS) đạt > 170, đảm bảo độ sắc nét và đa dạng.
- Hiệu suất suy luận: Tốc độ sinh ảnh nhanh gấp 1.5 – 2 lần nhờ tối ưu hóa lộ trình tự hồi quy và loại bỏ mã dư thừa.
- Tính nhất quán: Loại bỏ hoàn toàn lỗi đứt gãy cấu trúc; đảm bảo sự đồng nhất giữa bố cục tổng thể và chi tiết nhỏ (da, tóc, hoa văn).
- Đóng góp: Hoàn thiện bộ mã nguồn PyTorch và báo cáo phân tích Benchmark về mã hóa hình ảnh động.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1]. Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, Rianne van den Berg: Structured Denoising Diffusion Models in Discrete State-Spaces. NeurIPS 2021: 17981-17993

[2]. Hangbo Bao, Li Dong, Furu Wei: BEiT: BERT Pre-Training of Image Transformers. CoRR abs/2106.08254 (2021)

[3]. Emmanuel Bengio, Pierre-Luc Bacon, Joelle Pineau, Doina Precup: Conditional Computation in Neural Networks for Faster Models. CoRR abs/1511.06297 (2015)

[4]. Tolga Bolukbasi, Joseph Wang, Ofer Dekel, Venkatesh Saligrama: Adaptive Neural Networks for Efficient Inference. ICML 2017: 527-536